ERIA Working Paper
July 2012

Game-Based Assessment: An integrated model for capturing evidence of learning in play

Richard Halverson
Elizabeth Owen
Nathan Wills
University of Wisconsin-Madison

R. Benjamin Shapiro
Wisconsin Institutes for Discovery

Introduction

A central challenge for video gaming in education is to demonstrate evidence of player learning. A typical approach to assess learning in games is to measure the quality of player learning in terms of independent, pre-post instruments. This process can compare game-based learning against other kinds of interventions, but, in treating the game itself as a black box, we lose the unique characteristics of the games as a learning tool. James Gee has suggested that games themselves provide excellent models for designing the next generation of learning assessments. Well-designed games reward players for mastering required content and strategies, scaffold player activities toward greater complexity, engage players in organized social interaction toward shared goals, and provide feedback (through interface design) that allows players to monitor their own progress (Gee, 2007). Rather than ignore the motivating and information-rich features of games in capturing learning, assessment designers need to attend to the ways in which game-play itself can provide a powerful new form of assessment. This requires learning researchers to think of games as both intervention *and* assessment; and to develop methods for using the internal structures of games as paths for evidence generation to document learning.

This paper presents a Game-Based Assessment model (GBA) designed to capture data on player learning in the midst of game-play. The GBA model has been developed by the Games, Learning and Society (GLS) Research group as a process for capturing relevant information on play and testing whether it can constitute reliable evidence of learning. The GBA model draws on concepts and tools from evidence-centered design (e.g. Mislevy & Haertel, 2006), stealth assessment (Shute, 2011) and education data mining (e.g. Baker & Yacef, 2009) to describe a strategy for building assessment tools into game design from the ground up in order to use game play itself as the barometer of player learning.

This paper describes the GBA model and how it fits within the GLS game design process; provides an example of how the model works within a particular game (*Progenitor X*); and concludes with an account of what player interaction data tell us about learning in the game. The first section of the paper offers a brief overview of recent research in assessment and learning that provides the theoretical foundation for our argument. We then turn to a description of the GBA model as nested within the GLS game development process. We describe how GBA is grounded in the content model and game-flow design of the game development process, and then turn to the distinctive features of the GBA: the *semantic template* and *learning telemetry* layers. The semantic template allows researchers and designers to identify the key moments of player interaction in the midst of gameplay as hypotheses to capture significant information about player learning; learning telemetry provides a schematic for a generic information gathering mechanism to transform key-moment click-stream data into play profiles. The resulting profiles can then be matched to prior play records, other players, or pre-post measures to determine the validity and reliability of the GBA model, to identify which aspects of game-play provide salient data for learning. The next section details how the GBA was integrated into *Progenitor X*, a GLS game designed to teach processes of stem-cell science in the midst of a zombie invasion. The paper concludes with an analysis that demonstrates the value of peering inside the black box of game data. Our analysis shows that player learning is best predicted not by the number of times the game is played, nor by the number of successes or failures of the individual components. Rather, the best predictor of learning in the game is the *kind* of failure that players experience. As we

integrate GBA across other GLS games, we should be able to develop more robust hypotheses of which aspects of our games matter for learning, and how to better use games as both instruments and assessments for learning.

Game-Based Assessment Model

GBA begins with designing a game around specific learning goals. GBA has emerged as the design strategy of the Games, Learning and Society development group[1]. Our strategy is to bring content experts, game developers and programmers, artists, educators and learning scientists together in a collaborative design process (as described in Squire & Patterson, 2010). These design partners work to match subject matter content that can be best expressed in particular video game genres. GLS design teams identify promising content chunks that may enhance the public understanding of a particular domain. Typically these content chunks, such as the cultivation of a stem cell culture (*Progenitor X*)[2], the process through which a virus enters a cell (*Virulent*)[3], or how implicit bias influences perceptions in professional settings (*Fair Play*)[4] are laden with technical vocabulary and embedded in larger domain constructions. The task of the game design group is to translate core content chunks into games that invite players to participate in the logic of the concepts as a condition for learning the terminology. The development of a GBA model is critical for the design team to determine the relation between the game flow and the content model, in other words, to understand whether players can access the content chunks through game play.

Recent Research on Games, Learning, Assessment and Data.

The GBA design is grounded in recent research in game-based learning, evidence-centered design (ECD), and education data mining (EDM). We use a game-based learning experience to implement a version of ECD's task/content/evidence model into the game design. We then collect patterns of click-stream data, as in EDM, to develop records of in-game player interaction that can be used as evidence for learning. Here we briefly review some of the core research ideas that led to our GBA design.

Video games and learning. Kurt Squire asserts that "games differ from simulations in that they give roles, goals, and agency", and use "transgressive play" (Salen & Zimmerman, 2004) to "elicit fantasies" (Squire, 2011, p. 29). A key aspect of effective game design is the integration of data channels that inform both play and design. Gee (2005) highlights how good video games include just-in-time information (scaffolding) and cycles of expertise. Good games include formative assessment cycles that foster ongoing feedback and customize player difficulty levels (Shute, 2011). In order to maintain this immersive context for learning, good games consist of ongoing assessment balanced with engaging mechanics and narrative (Squire, 2006). Good games are not only scaffolded, engaging designed experiences (Squire, 2006), they also hold the power to improve learning. Situated learning theory suggests that learning exists *in situ*, inseparable from environment or context (Brown & Collins, 1989). Virtual game worlds have been shown to provide a powerful environment for learning, supporting apprenticeship and collective higher-order thinking skills

[1] http://www.eriainteractive.com/index.php
[2] http://www.eriainteractive.com/project_ProgenitorX.php
[3] http://itunes.apple.com/au/app/virulent/id438485177?mt=8
[4] http://www.eriainteractive.com/project_Pathfinder.php

(Steinkuehler & Duncan, 2008). The communities of practice (Lave & Wenger, 1991) that emerge around games (Steinkuehler, 2006) amplify in-game learning by fostering collective intelligence and participatory cultures (Jenkins, 2006).

Games are also engaging. Research across several disciples suggests that interest sparks learning. The affective filter (Krashen, 1985) is an impediment to learning caused by negative emotional responses to instructional input – an impediment that can be lifted through permissive, engaging learning environments in which the student feels a sense of agency. Teachers relate how engaging lesson plans that give learners agency often result in better outcomes and fewer classroom management problems. This idea of personal empowerment is also central to the immersive "projected identity" of gameplay (Gee, 2003). Compelling, pleasurable learning experiences (Gee, 2007), videogames often facilitate interest-driven learning (Squire, 2011). Recent literacy & gaming studies have shown that interest-driven learning context can significantly impact student self-efficacy (Owen et. al, 2012), and result in dramatic increases in reading comprehension (Martin & Ochsner, forthcoming).

Thinking about assessment in gaming often leads researchers and designers outside of games to objective, non-game-based measures of successful learning. Game-based learning challenged this separation of assessment from learning. Games not only provide data-rich environments grounded in the best ideas of formative assessment, they also provide engaging environments in which players seek to develop skills by exploring complex narrative worlds. GBA plays on this data-richness and advantages of the game space to use game mechanics and narrative as a *foundation* in which to seamlessly embed assessment.

Evidence Centered Design. ECD is a leading model for assessing knowledge and skills in complex domains that "enables the estimation of students' competency levels and further provides evidence supporting claims" about the knowledge and skills being assessed (Shute, 2011, p. 508). According to John Behrens, ECD "is flexible enough to accommodate the affordances of new technologies and the demand to measure new domains while providing a united framework to describe current practice across a wide range of assessment activities" (Behrens et. al., 2012, p. 47). This broad applicability gives ECD a universal appeal, while leaving open the opportunity to develop other ECD-inspired assessment models specifically tailored to individual learning technologies.

ECD includes three key layers: 1) a *competency model* (CM) that defines key knowledge and skills to be assessed; 2) an *evidence model* detailing what behaviors or performances should reveal the CM's constructs; and 3) a *task model* that specific certain activities to elicit the behaviors that comprise evidence (Shute, 2011). ECD seems to work well with simulations than can be built in terms of the specified task model (e.g. Mislevy, 2011). However, adapting ECD to vide games has proven challenging. Val Shute notes that "making valid inferences about what the student knows, believes, and can do without disrupting the flow of the game" is a "main challenge" of educators in using games to support learning (Shute, 2011, p. 508). Assessing in-game performance is a "complex process that needs to take into account not only the engaging or motivational aspects of the activity but also the quality criteria that are needed according to the type of assessment that is being developed" (Zapata-Rivera & Bauer, 2012, p. 149).

Stealth assessment (SA) is a recent ECD-based approach to "identify key competencies and use games as instructional learning vehicles" (Shute, 2011, p. 505). SA, in essence, is an ECD-based model that is focused on connecting 21st century skill competency models to existing video games. The competency model in stealth assessment is framed in terms of 21st century skill (like Creative Problem Solving), which can then be documented with click-stream data and analyzed with Bayesian network techniques (Shute, 2011). An evidence model based on the CM is created, and then aligned with a task model (which defines player action within existing game mechanics). Val Shute's work explores how stealth assessment can link game-play goals and desired skills and knowledge. Her work matches the content model goals in the game (e.g. collaborate with other players to slay dragons) with competency model skills (e.g. 21st century skills). Recent efforts to design stealth assessment models have focused more on teaming the game and assessment designers so that the content and task models can be better aligned. GBA follows up on this insight by designing games around content models to create a seamless assessment/play experience.

Educational data mining. Educational Data Mining's (EDM) approach to assessment is to explore the click-stream data that result from participation in virtual worlds for patterns of user interaction, and, hopefully, evidence for learning. EDM is "an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings" (http://www.educationaldatamining.org/). EDM seeks to pull data from "interactive learning environments, computer-supported collaborative learning, or administrative" databases and analyze it according to "properties in the data itself, rather than in advance" (http://www.educationaldatamining.org/). Data modeling techniques (e.g. statistical cluster and factor analysis) are used to identify patterns that allow analysts to make inferences about learning outcomes (Xu & Recker, 2010).

EDM is agnostic about a pre-defined competency model to guide data collection. Much of the early EDM work has taken place to analyze the kinds of mistakes that students make when interacting with cognitive tutors. (Baker, Corbett et. al. 2010). Identifying patterns of user interaction allows designers to build computer-adaptive learning models that can anticipate the kinds of problems that will challenge the current student skill levels. The GBA model seeks to build this kind of system-adaptivity into the kinds of games situated in worlds far more open than the typical tutoring system. We hope to build on the ECM model of event-stream telemetric data collection in learning games is a promising new frontier in the world of educational research.

ECD, SA, and EDM have provided a strong precedent of assessment research in digital environments. The competency-based, flexible assessment framework of ECD - and by extension SA's gameflow-based approach - paves the way for innovative modeling of assessment in educational game design. EDM's click-stream data mining in learning spaces invites a natural extension into event-steam telemetry in educational game worlds. Together, they open opportunity for a new game-specific assessment models integrating ground-up game design with embedded assessment mechanics, based on a common content model and aligning event-stream telemetry to draw inferences about learning.

The Conceptual Model of GBA.

The GBA model is designed to draw significant game-play moves from the game-context. The model has is integrated into an overall 4-layer GLS game design strategy: the *content-model*; the *game flow* design; and the GBA components of the *semantic template* and *learning telemetry (Figure 1)*. The first two layers, the content model and the game flow design, constitute the game design process. The content model outlines the learning goals for the game. The game flow design builds player interaction opportunities around these learning goals to create a gaming experience. The final two layers, the semantic template and the learning telemetry, form the assessment process. The semantic template selects relevant data from the click-stream generated by game-play; the learning telemetry layer collects and organizes the resulting data-record into player-profiles. Here we provide a brief overview of how these layers, taken together, comprise a generic blueprint for our approach to assessment-driven game design.

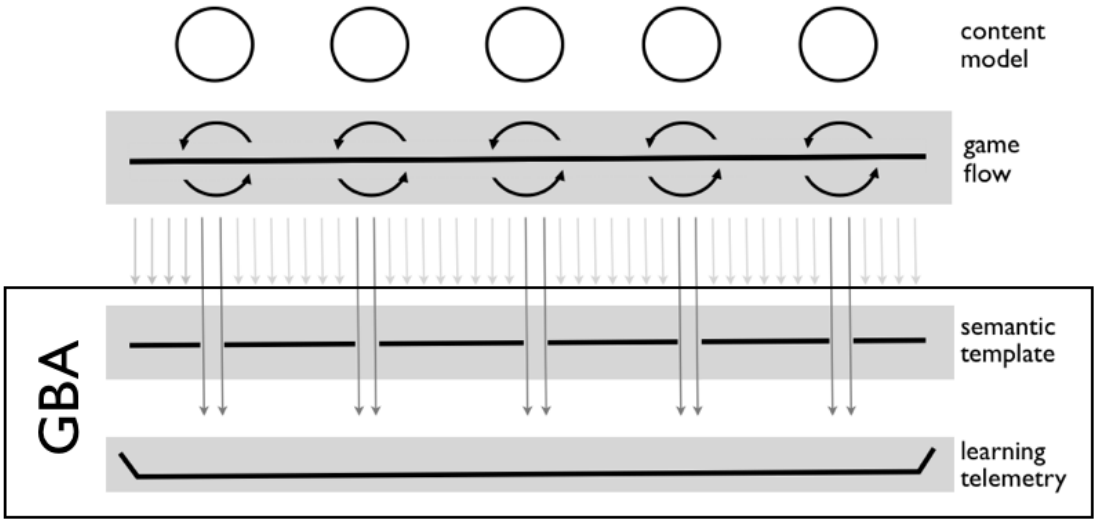


Figure 1: GLS Game Design Layers

Content Model. The content model for a GLS game consists of several content chunks along a process that represents current thinking and practice and in given domain. As with evidence-centered design, the content models are selected to highlight an important aspect of a complex domain. However, rather than aiming to represent abstracted expertise (e.g. 21st century skills), the content model chosen for GLS game design tends toward much more modest selections of content ideas that both experts and the designers feel would resonate with a general lay audience. Because the resulting medium for interaction is a game, rather than a simulation, the design team is concerned with creating motivating conditions of play as well as the representational accuracy of the content model.

The content model for the *Progenitor X* game provides an example. The *Progenitor X* project began as a collaboration between GLS designers and a UW-Madison stem-cell lab. After several discussions about the scope of a potential game (including proposals to address public misunderstandings about the sources and uses of stem-cells, and the political controversies around stem cell policy), the team decided to focus on the lab processes by which scientists manipulate stem cells and use them in treatment. The team felt that these

basic science questions were often overlooked in the public debate, and that game-based experience with the science of stem cells could enhance the public understanding of the science as well as to encourage younger players to consider this kind of research as a career option.

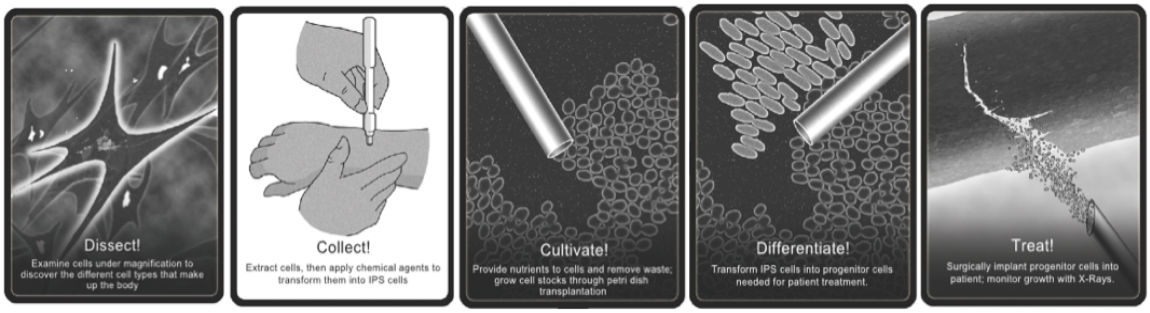


Figure 2: *Progenitor X* Content-Model

The content model invited players to dissect, collect, cultivate, differentiate and treat diseased tissues with stem cells (*Figure 2*). Each verb in the content model provides an occasion for interaction. A process derived from a professional practice provides a thread that links the concepts together into a simplistic but coherent account of the process that actually guides the work of scientists and doctors. The game design process for each of the GLS games (as well as most other learning games) begins with a similar, simplified content model to be translated into a game genre.

Game-flow design. A GLS game is designed as to motivate player interaction and as the occasion for player learning. Achieving both of these tasks is important for the success of a learning game. Through the iterative design process, the content model is embedded in an interactive world that allows players to interact with the core ideas. The verbs of the content model are translated into "key moments" (Halverson & Gibbons, 2010) in game play in which players make decisions about their progress. In this sense, the game is the ECD task model, but with the added necessity of providing a compelling context for players to interact with the content model. In other words, the game-flow design seeks to create player experience to interact with domain content while navigating the norms, roles and the narrative structure of a simulated world.

The game-flow design is named to emphasize how the game provides a designed experience (Squire, 2006) for the player, rather than a straightforward assessment of content model mastery. This requires situating key content-model verbs into critical choices of game play. These critical game-play choices must reflect the dual requirements of engaging players in the game-narrative and authentically interacting with core content. The conventions of, for example, a single-player shooter or a real-time strategy game, provide occasions for the repetition of key concepts in ways that support drill-based learning. Integrating these genre conventions into a compelling narrative allows for game designers to build cycles of repetition and exploration into game-flow. In the final section of the paper, we will provide an example of how (and whether) *Progenitor X* constructs skill building and adventure gaming into a coherent experience.

Semantic template. The semantic template and the learning telemetry layers form the main components of the GBA designed to select and store certain key indicators of player interaction. Game-worlds generate an enormous amount of player interaction data. This data richness provides an important advantage of virtual environments (when compared to real-world learning environments) for tracking player interaction and performance. However, the click stream data that results from player interaction is typically too rich, and too disconnected from the game-flow, to support direct inferences about player interaction, much less player learning. The semantic template is designed to "narrow" the data that result from play in ways that allow designers to make inferences about how players interact with key game-flow episodes, and to compare records of player interaction with the content model.

The semantic template is built in coordination with the game-flow design. The key question for semantic template design is: Of all the clicks that players make in the game, which ones indicate learning? The semantic template represents a hypothesis about which in-game actions can generate interesting evidence of learning. These hypotheses can then be confirmed by other analyses and used to inform subsequent game design. In some ways, the construction of a semantic template is similar to the design of a scoring system. Game scores represent designed markers of player progress. The point system of a game like *Diablo III*, for example, shows how designers use in-game achievements to help players understand their progress. Points are rewarded for completing quests and battles, are used to promote players to subsequent levels, which in turn opens new quests and battles. A semantic template provides a slight twist to a scoring model of player progress. In addition to tracking in-game progress by recording play achievements, the semantic template also records learning progress by recording how players interact with the content model.

Learning telemetry. The learning telemetry layer collects the data specified by the semantic template and organizes it for analysis. It is a mechanism of the game environment that coordinates the different components of the game world into a sequential data-stream that enables analysts to track player paths across the game-world. Telemetry systems are already widely used in game-worlds and digital environments to collect data on player/user interaction (Gagne & Seif El-Nasr, in press). Typically, the data that result from telemetry analysis are used for detecting game bugs (Niwinski & Randall, 2010), improving player experience (Dankoff, 2011) or for connecting users to advertising (through sites such as Facebook and Google+). GBA adapts the telemetry concept to focus on collecting data specific to learning so that we can gauge player interaction with the content level and improve game-flow design.

The semantic template and learning telemetry layers, taken together, constitute the GBA model. Data on player interaction are generated from the game-flow system. The semantic template determines which data to track, and the learning telemetry layer collects the resulting data for analysis. The resulting profile of player interaction opens up the black box of game play to analyze the relationship between play and learning. Analysts can, for example, compare play in-game play profiles to pre- and post-tests to determine the relations between player interaction and content learning. Player profiles can be compared to one another as well, to determine the types of player interaction (e.g. the pace, the time spent in each area, the degree to which players explore the game space or use learning resources) and their relation to learning goals. Player interaction data can also be dynamically visualized to

trace (and compare) patterns of player interaction with the system. These forms of analysis will allow designers to understand which aspects of the game are correlated with learning gains, and point designers toward the aspects of the game that can be tweaked. Reliable evidence about in-game learning can also help create adaptive gaming environments, in which players with demonstrated competencies can be given challenges that invite demonstration of mastered skills and knowledge, or quests that take a different tack on content goals players had difficulty mastering. These kinds of records-of-play may help bring the techniques of computer tutoring environments (e.g. Koedinger & Corbett, 2006) to bear in game-play spaces.

Of course, presenting a conceptual model for game-based assessment design is not the same as showing that the model can actually be implemented, or that it can generate reliable insights about play, learning and design. We are not yet in a position to deliver results that would confirm the viability or the quality of the GBA model across game spaces. However, in the following section, we will show how we have built the central ideas of the GBA into *Progenitor X*, a game designed to teach about stem-cell science in a world overrun by ravenous zombies.

Building GBA: Progenitor X

Progenitor X is a narrative-based, turn-based game involving a series of puzzles designed to teach basic practices of stem-cell science. The game was developed in partnership with the Regenerative Medicine[5] research team to present core ideas of cutting edge sciences in context of a game. *Progenitor X* players are challenged to cultivate and differentiate stem cells, assemble tissues and replace organs that have been contaminated with a zombie virus.

Figure 3: *Progenitor X* Cell Level

Completing game play requires players to solve 10-12 cell, tissue and organ puzzle cycles. Players initially encounter a cell cycle that involves a sequence of treatment and collection tools that transform pluripotent stem cells into particular cell types (Figure 3). Tissue cycles require players to layer successfully transformed cells into segments of tissue; the organ level

[5] http://stemcells.wisc.edu/faculty/thomson.html

requires the assembly of tissue segments into organ shapes (Figures 4 & 5). While players learn the cell cycle first, subsequent play requires players to repeat cell-tissue, then cell-cell-tissue-cell cycles in order to complete the game. The final cycle of the game, the organ cycle, functioned as a boss-level that required players to use all the skills learned in the cell and tissue cycle (e.g. a cell-cell-tissue-cell-tissue cycle sequence) to complete the game.

Figure 4: *Progenitor X* Tissue Level

Figure 5: *Progenitor X* Organ Level

Assembling the GBA required designers to build a system that would collect data on player interaction with the system, then to select which key moves/clicks in the game might correlate with content learning gains. The first stage of GBA development involved building the semantic template from stages of game play. We identified 15 key moments in game

play, typically occasions in which players are asked to use tools in new ways or in combination to transition to the next cycle of play, that might yield important user data. Constructing the semantic template required the design team to specify the mission and the cycle key moments. It also led the designers to articulate three modes of aspects interaction within the key moments: the type of cell manipulated (Stage 1); the treatment tool used (Stage 2); and the type of cell collected (Stage 3). (Figure 6 provides some detail from the *Progenitor X* semantic template.). These key moment events constituted our hypotheses about where we would be able to locate evidence for learning in the game-flow.

In-Game Sequence	Mission	Cycle**	Cell population type (Stage 1)	Treatment tool type (Stage 2)	Collection cell type (Stage 3)
1	1	A. ips	fibroblasts	electroporate	ips
2	1	B. meso i	ips	growth factor	meso
3	1	A. ips ii / C. ecto i	fibroblasts	electroporate	ips
4	1	C. ecto ii	ips	growth factor	ecto
5	2	E. tissue i	meso	N/A	meso / tissue
6	2	E. tissue ii	meso	N/A	meso / tissue
7	2	A. ips iii / D. endo i	fibroblasts	electroporate	ips
8	2	D. endo ii	ips	growth factor	endo
9	2	E. tissue iii	endo	N/A	endo / tissue
10	3	F. organ i	N/A	scan	necrotic tissue
11	3	E. tissue iv	meso	N/A	meso / tissue
12	3	F. organ iii	N/A	scan	necrotic tissue
13	3	A. ips iv / B. meso ii	fibroblasts	electroporate	ips
14	3	B. meso iii	ips	growth factor	meso
15	3	E. tissue v	meso	N/A	meso / tissue
		**(A = ips, B-D = diff cells, E = tissue, F = organs)			

Figure 6: *Progenitor X* Semantic Template

The next challenge was to locate the key moments of the semantic template within the data-stream generated by play. The click-stream data needed to be shaped by triangulating player moves and game-events into a coherent time-line that could parallel to game play flow. The *Progenitor X* design team developed this learning telemetry layer by tagging each move players made in the game with information such as the button used, a time-stamp, whether the action suggested by the system was flashing, the name of the tool used in the action, and type of information included in the tag (Figure 7).

▼ _id	4fbd2a48cf86304117000004	ObjectId
_id	4fbd2a48cf86304117000004	ObjectId
created_at	2012-05-23 13:19:52 -0500	Date
gameName	ProgenitorX	String
key	ToolSelectedData	String
schema	4-18-2012	String
session_token	2012-05-23_115607	String
timestamp	854	Int
toolName	Move	String
updated_at	2012-05-23 13:19:52 -0500	Date
user_id	235	Int
wasFlashing	YES	Bool

Figure 7: Sample Learning Telemetry Data

Once events were tagged, the resulting telemetry data stream allowed the designers to query the data stream in order to identify key moments in the context of game flow. The telemetry level recreated a sense of "played time" through a series of rich data points that allowed the design team to map the semantic template onto the data stream.

Building the GBA involved working the content experts and game designers to identify the semantic template key moments in game play that might elicit evidence of learning from play; tagging the data stream to recreate a sense of data-flow in a learning telemetry layer that could parallel the experienced game-flow; and developing query tools that could map the key moments of the semantic template onto the data-stream captured by the learning telemetry. Our adaptation of telemetry techniques to capture learning, rather than, for example, debugging or marketing purposes, indicates our attention to the correspondence of game-play and the underlying content level. The GBA concepts and architecture open the way to developing generic tools for key moment recognition and data-stream recreation that will be used across GLS games, and, hopefully, extended to the development of any future games for learning.

GBA Analysis: The Role of Far Failure.

Our main question for analysis was whether (and how) in-game player action related to content learning outcomes. In our experience, many game-based researchers hope for a correlation between successful play and positive learning outcomes. Our implementation of the GBA allowed us to look more deeply into the meaning and patterns of "successful play" and to draw more nuanced conclusions about the relations between play and learning.

The GBA identified key moments within game play and captured data aligned with the semantic template to yield a very rich set of information about player action in *Progenitor X*. The analysis was conducted on data collected from 39 middle and high school students who played *Progenitor X* over a three-week period. In addition to playing through *Progenitor X*, each student completed a brief assessment before and after play to provide an independent measure of content knowledge that we could use to compare with game-play patterns (Appendix A). The pre-post tests indicated that players learned stem-cell science content as a result of play. Players had a statistically significant increase in scores (11% gain, $p=.01$). This significant increase in content understanding after gameplay gave a rich comparison point for analyzing specific movements during the game.

We constructed within-player and across-player patterns in the GBA data, and compared the results with the pre-post assessment to begin to look into the black box of game-play data. Our analysis discussion includes three main sections: first, we discuss some of the patterns in comparing aggregated player data with the pre-post assessments; second, we investigate some of the patterns that emerged within the game-play cycles across players; third, we consider within-player differences in the kinds of failure experienced in the game, and how these differences relate to the pre-post learning measures.

Aggregate trends

Our first analysis focused on game data aggregated across players. In order to sort the player data into meaningful patterns, we developed an efficiency ratio that measured the number of successful cycle completions by a player over the number of times the cycle was tried. For example, if a player successfully collected the required number of cells in a cycle 2 times, and tried to complete the cycle 5 times, the player's efficiency ratio would be 40%.

$$\text{Efficiency Ratio} = \#\text{ of successes}/\#\text{ of tries}$$

The higher the percentage, the more efficient the play. The efficiency ratio operationalizes our assumption that mastery of the game mechanics means the improving successful cycle navigation with fewer tries. Our initial hypothesis, then, was that improved play efficiency would be an indicator of successful learning, both of the game mechanics and the underlying content model.

	Game Progress	**Pre-Post Gains**
Total Gameplay		11% average increase (t-test sig = .0098)
Efficiency Ratio	significant positive correlation* (r = .3254)	no significant correlation
Boss-level Efficiency Ratio		significant positive correlation (r = .3219)

n=39, α = .10

Table 1: Aggregate *Progenitor X* Data Summary

Our findings on this initial hypothesis were mixed. Player progress through the game was positively correlated with efficiency ratio. In other words, players with higher efficiency ratios had higher game completion rates. However, the aggregate efficiency ratios told us little about learning outcomes as measured by the post-test. Neither the total numbers of tries, the total numbers of successes, nor the aggregate efficiency ratio (across cycles) were significantly correlated with pre-post learning gains. Only in the last cycle of the game (the organ cycle boss level) was the efficiency ratio correlated with pre-post gains (r = .3219). Thus, by the end mission of the game, being good at the mechanics was associated with learning the content model. However, we were unable to identify game mastery (as measured by player efficiency ratio) with content learning (as measured by pre-post tests) from aggregate game-play data. This led us to investigate what was going on with players within the specific game cycles.

Cycle-Specific Examination

Our next step was to investigate patterns of player interaction data within the individual game cycles. We found two indicators in which interaction data were positively correlated with the pre-post learning gains (Table 2). Both indicators occurred in the cell cycle of play.

	Game Progress
# of cell cycle starts	Significant negative correlation* (r = -.2965)
# of cell cycle destroys	Significant negative correlation (r = -.3390)
# of TOTAL cycle starts (all types)	No significant correlation

n=39, α = .10

Table 2: Progenitor X Cycle Starts and Destroys

The first indicator was the number of times a player started a cell cycle; the second was the number of time the cell cycle was destroyed (i.e. the player failed to collect the appropriate number and kind of cells). In both cases, the number of cell cycle starts and the number of cell cycle were negatively correlated with pre-post learning outcomes, while the total starts and destroys across the game were not significantly related to learning. What was going on for players in the cell cycles?

In order to examine player interaction action, we mapped all possible outcomes for the cell cycle play. Play in the cell cycle requires players to populate an initial grid with the kinds of ips cells that can be transformed into the target population, and to manipulate the cells toward the appropriate transformation. There were four ways that cell cycle play could end (Table 8). Player can end play in in one successful path, by populating the initial grid with the right cells, and collecting the correct kinds of cell. Players can fail in three ways. First, they can conduct the initial population correctly, but take too many turns manipulating the cells, which causes the Ph in the culture to become toxic (the equivalent of having the health meter run out). Second, they can conduct the initial population correctly, but fail by collecting the wrong kinds of cells (incorrect cell collection). Third, they can incorrectly populate the cell with the right kinds of cells at the outset, and end a cycle by collecting incorrect cells while the health meter runs out.

Action	Correct Collect	Ph Meter Fail	Incorrect Collect	Ph Fail + Incorrect
Initial Population?	✓	✓	✓	
Completed in time?	✓		✓	
Success?	✓			

Table 3: Progenitor X Cell Cycle Outcomes

The possible player outcomes imply varying degrees of player compliance with in-game cues (e.g. flashing buttons, in-game narration, suggested play examples). The redundance of in-game cues suggests differences between at least two kinds of failure in the game: players who complied with game cues but failed to master the mechanics, and players who did not heed the game cues and tried to "game" the mechanics. Our second hypothesis was that players who followed the cues, but simply ran out of time, had a better chance of learning the content model when compared to players who disregarded the game's guidance system.

Figure 8: Progenitor Cell Cycle Player Outcomes

To explore this idea, we grouped the latter two types of failures into "near" and "far failure." (Figure 8). We thus grouped 3 possible player outcomes: A) correct collection (successful) ; B) correct set-up but health runs out (near failure) and C) incorrect collection and/or time runs out (far failure).

The analysis of far failures gave considerable insight into the player data. While the total number of failures (of all kinds) had no relation to cell cycle completion, the number of far

failures across players was negatively correlated with cell cycle completion. We also found that the number of "far failures" for players *across all cycles* was negatively correlated with learning as measured by the pre-post tests (r = -.2788; **p=?**). Other indicators of play, including the number of cycles started, number of successful collects, and total number of cycles completed had no correlation with pre-post gains.

To deepen our understanding of far failure, we compared the incidence of these types of failure in the player groups in the upper and lower quartiles of pre-post assessment gain. The upper player quartile included 10 students who averaged a 33% increase in content model learning; the lower player quartile included the same number of students who averaged a 17% decrease in the pre-post learning score. With the upper quartile players, the number of cell cycles started (and the number of times the grid was destroyed) was positively correlated with learning gains. The lower quartile was opposite: cell cycles started (and number of grid destructions) were *negatively* correlated with learning gains. This implied that the upper quartile was learning more than the lower quartile during each cell cycle, whether they failed or succeeded.

The lens of far failure uncovered further differences in quartile group comparisons (Table 4). On average, students in the lower quartile had 7 cycles of far failure, while upper quartile students only had 2.3 cycles. Since both groups had comparable total NUMBERS of failures, the lower quartile had a greater proportion of far failures; thus, their losses in learning the content may be linked to the quality of their responsive to the game queues. The pre-post correlation with this number suggests that certain types of failure, not failure itself, inform learning.

	Upper Q	Lower Q
# cell starts	Sig. + correlation with pre-post (r = .7059)	Sig. - correlation with pre-post (r = -.8641)
# of cell destroys	+ correlation* with pre-post (r = .3401)	Sig. - correlation with pre-post (r = -.5261)
# of "far" failure cycles (average)	**2.3**	**7**
	Sig. + correlation with pre-post (r = .5796)	Sig. - correlation with pre-post (r = -.9408)
Total failures	No significant difference.	No significant difference.

* n=20, α = .10

Table 4: Progenitor X Far Failure Analysis

Discussion

The GBA model allowed us to move beyond a simple pre-post comparison of game play to learning outcomes by providing data on how players interacted with the game environment. The design of the semantic template allowed us to collect data at (what we believed to be) key moments in the course of game-play; the learning telemetry allowed us to tag and assemble these click-stream data points into play profiles we could use for analysis. The GBA model allowed us to blend the structural aspects of evidence-centered design with the openness to player click-stream data of education data modeling to generate information flows that informed our understanding of game-play and learning outcomes.

Our initial findings that game completion led to pre-post gains led us to explore whether game-play data could deepen our understanding of learning. We found that completion of the organ boss level, in which players recapitulated prior cell and tissue levels, was correlated with the pre-post games. However, we found no significant correlation with the number of tries, the number of failures, or the number of times players played the game and the learning outcomes. To further explore the player data, we distinguished between the ways in which players could fail in the game. We found that two kinds of failure, the far failure condition, were negatively correlated with pre-post learning gains. Players with the highest learning gains experienced 2.3 far failures in the course of game play, while players in the lowest quartile experienced 7 far failures.

What do these data mean in the context of game play? A common challenge of games for learning research is the difficulty of disentangling mastery of the learning mechanics from learning outcomes. Playing the game, in other words, is not the same as learning the underlying content model (e.g. Clark & Martinez-Garza, in press). Much of the connection (or disconnection) between the game-flow and the content model rests on the quality of the game design. Good games translate the content model to the player through compelling in-game moves and strategies, while poor games allow players to mash buttons and "game" the environment while bypassing the content model. Far failure is another way of describing a play-style of "gaming" the *Progenitor X* environment. Players who click cells and tools not in compliance with the content model were notably less successful in learning the underlying content than players who attended to the in-game prompts. This suggests that the Progenitor X game mechanics were attuned to the content model, and that playing the game as designed allowed players to successfully learn the content.

The data provided by the GBA model were able to provide some insight into the role of failure in *Progenitor X* game play. Games allow players to experiment with failure without real-world consequences. However, the kinds of failures players experience matter. Productive failure (Kapur, 2008) suggests that effective learning environments encourage students to activate prior knowledge as a condition for direct instruction. *Progenitor X* introduces players into an unfamiliar subject matter context (regenerative medicine), but in a familiar game-genre context (puzzle-based video games). Familiarity with the game-conventions invites players to interact with a system in order to learn the programmed relationships between cells, tissues, tools and cultures. One way to interpret the results of our analysis is that productive failure happens when players bridge game-mechanic knowledge to content-model knowledge through game-play; non-productive failure happens when players ignore the content model and treat *Progenitor X* as a colorful puzzle game with zombies. The richness of the data generated by the GBA will allow us to further explore the relations between player interaction and learning.

In addition to informing our understanding of player learning, the GBA design also generated valuable information on game redesign. The design team was able to use the semantic template design, for example, as a tool to clarify how players should interact with the game cycles. In one cycle, players were asked to prepare a report on the correct configuration of cells and tissues for recreating zombie-resistant hearts. The initial design of the report was more of a perfunctory click-though screen that allowed players to proceed to the next challenge in the game. GBA designers saw the report development as a valuable occasion for players for reflection-on-practice that would summarize in-game experience for

communication to a non-playing character. Building in-game consequences for the quality of the report (e.g. making key resources available based on report recommendations) suggested game mechanic ideas that would embed reflective practices in the context of game-flow.

The design concepts and data we have presented here also highlight several limitations in the scope and implementation of GBA. First, the "bet-hedging" role the semantic template plays may narrowing the data too severely and can leave out the very information necessary to understanding player interaction and learning. A key insight of education data mining research is to include as much of the data as possible on the grounds that the analyst never knows which patterns of game-flow moves might be most important.

Second, our discussion of the GBA overlooks the advantage that game-based learning has over other learning interventions. James Gee (2008) describes how the potential of games for learning means that "little g" games (such as *Progenitor X*) should be nested in "Big G" game contexts that activate play-based learning in social and knowledge rich interaction contexts. By defining learning in terms of understanding the content model, we deliberately focused attention of the GBA design on the little g game outcomes, and have deliberately avoided mention of the larger contexts, such as the narrative context of zombie games, or the social contexts of interactive play, that bring game-based learning alive. Similarly, by focusing on how games provide access to a particular content model, our presentation of the GBA has ignored the emergent (Steinkuehler, 2004) or transgressive (Aarspeth, 2007; Kafai, Fields & Giang, 2009) ways that players transform the game experience and outcomes through play. We feel that narrowly defining the GBA to omit what some scholars claim to be the central contributions of games to learning is a serious issue with our presentation of the GBA.

Our response to these comments is to emphasize the modest, preliminary nature of our investigations. The guiding question of the GBA design was not to issue definitive statements about the future of game-based assessment, rather, it was to address the much more modest goal of whether we can say anything interesting about learning with structured in-game data. The semantic template allowed us to build "mid-level" hypotheses about which data might count for learning, and led us to analyses that correlated the relation between patterns in these data points with pre-post tests. While more sophisticated data-mining techniques will enable future iterations of GBA research to explore nuanced patterns of interaction and learning, we feel that the insertion of a semantic template between the game-flow and the telemetry layers provides a model of "theory-driven" inquiry that can supplement, rather than supplant, the more "brute-force" search techniques of data-mining.

Similarly, the modest aims of GBA to say something interesting about learning is intended to supplement, not supplant, the more ambitious agenda of locating little g learning in Big G contexts. If the little g game can generate reliable information on content learning, then embedding GBA architectures in a Big G context may allow for a much wider set of comparisons within and across players and contexts, thus enriching, rather than diminishing, the range of inquiry open to game-based learning researchers. Our goal is to integrate some version of GBA architecture across GLS learning games, then, in collaboration with our partners, linking data flows from game environments across play contexts and even into school information systems. Our agenda is to use GBA to show that games can serve as assessments that generate reliable evidence for content learning to formal education

environments, which can then legitimate the liberating potential of games as the ultimate disruptive technology to shake and rattle the social conventions that limit the potential of learning technologies in schools.

References

Aarseth, Espen (2007) "I Fought the Law: Transgressive Play and the Implied Player". Proceedings of DiGRA 2005 Conference. http://www.digra.org/dl/db/07313.03489.pdf.

Baker, R.S.J.d., Corbett, A.T., Gowda, S.M., Wagner, A.Z., MacLaren, B.M., Kauffman, L.R., Mitchell, A.P., Giguere, S. (2010) Contextual Slip and Prediction of Student Performance After Use of an Intelligent Tutor. *Proceedings of the 18th Annual Conference on User Modeling, Adaptation, and Personalization*, 52-63..

Baker, R. & Yacef, K. (2009) The state of educational data mining: A review and future visions. *Journal of Education Data Mining 1*(1) 3-17.

Behrens, J. T., Mislevy, R. J., DiCerbo, K. E., & Levy. R. (2012) 21st Century Dynamic Assessment, In Mayrath, M, Clarke-Midura, J., Robinson, D. and Shraw, G. *Technology-Based Assessments for 21st Century Skills: Theoretical and Practical Implications from Modern Research*. Pages 42-53. Information Age Publishing.

Brown, J. S., Collins, A., & Holum, A. (1991). Cognitive Apprenticeship: Making thinking visible. *American Educator*, 6-91.

Clark, D. B., & Martinez-Garza, M. (in press). Prediction and explanation as design mechanics in conceptually-integrated digital games to help players articulate the tacit understandings they build through gameplay. C. Steinkuhler, K. Squire, & S. Barab (Eds.), *Games, learning, and society: Learning and meaning in the digital age*. Cambridge: Cambridge University Press.

Dankoff, J. (2011). Game Telemetry with Playtest DNA on Assassin's Creed. Engine Room.ubi.com blog site. Accessed at http://engineroom.ubi.com/game-telemetry-with-playtest-dna-on-assassins

Gagne, A. and **Seif El-Nasr, M.** (submitted). Analysis of Telemetry Data from a Real Time Strategy Game: A Case Study. *ACM Computers in Entertainment.*

Gee, J. P. (2003). What Video Games Have to Teach Us About Learning and Literacy. New York: Palgrave Macmillan.

Gee, J. P. (2005). Learning by Design: good video games as learning machines. *E-Learning* **2** (1): 5–16.

Gee, James Paul. (2008) "Learning and Games." *The Ecology of Games: Connecting Youth, Games, and Learning.* Edited by Katie Salen. The John D. and Catherine T. MacArthur Foundation Series on Digital Media and Learning. Cambridge, MA: The MIT Press, 2008. 21–40. doi: 10.1162/dmal.9780262693646.021

Halverson, E. R., and Gibbons, D. (2010). Key Moments' as Pedagogical Windows into the Digital Video Production Process. *Journal of Computing in Teacher Education 26(2): 69-74.*

Jenkins, H., Purushotma, R., Clinton, K., Weigel, M., & Robison, A. (2007). *Confronting the challenges of participatory culture: Media education for the 21st century.* MacArthur Foundation Digital Literacy Series. Available at http://newmedialiteracies.org/files/working /NMLWhitePaper.pdf

Kafai, Y. B., Fields, D. & Michael T. Giang. Transgressive Gender Play: Profiles and Portraits of Girl Players in a Tween Virtual World. Breaking New Ground: Innovation in Games, Play, Practice and Theory. Proceedings of DiGRA 2009

Kapur, M. (2008). Productive failure. *Cognition and Instruction, 26*(3), 379-424.

Koedinger, K. R. & Corbett, A. T. (2006). Cognitive Tutors: Technology bringing learning science to the classroom. In K. Sawyer (Ed.), *The Cambridge Handbook of the Learning Sciences,* (pp. 61–78). Cambridge, UK: Cambridge University Press.

Krashen, S.D. (1985), *The Input Hypothesis: Issues and Implications*, New York: Longman

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation.* Cambridge, UK: Cambridge University Press.

Mislevy, R. J. & Haertel, G. D. (2006). Implications of Evidence-Centered Design for Educational Testing. Principled Assessment Designs for Inquiry Technical Report 17. SRI International Center for Technology in Learning. Accessed May 31, 2012 at http://padi.sri.com/downloads/TR17_EMIP.pdf

Niwinski, T., Randall, D.J. (2010) Using Telemetry to Improve Zombie Killing. GDC Canada, May 2010. Accessed at http://www.gdcvault.com/play/1013152/Using-Telemetry-to-Improve-Zombie

Salen, K., & Zimmerman, E. (2003*). Rules of Play: Game Design Fundamentals*. The MIT Press.

Shute, V. (2011). Stealth assessment in computer-based games to support learning. In S. Tobias & J.D. Fletcher (Eds.) *Computer Games and Instruction.* Information Age: Charlotte, NC. 503-523.

Squire, K.D. (2006) From content to context: Video games as designed experience. *Educational Researcher* 35(8) pp. 19-29.

Squire, K. D. (2011) *Video Games and Learning: Teaching and Participatory Culture in the Digital Age*. Teachers College Press: New York.

Squire, K & Patterson, N. (2010) Games and Simulations in Informal Science Education. WCER Working Paper No. 2010-13

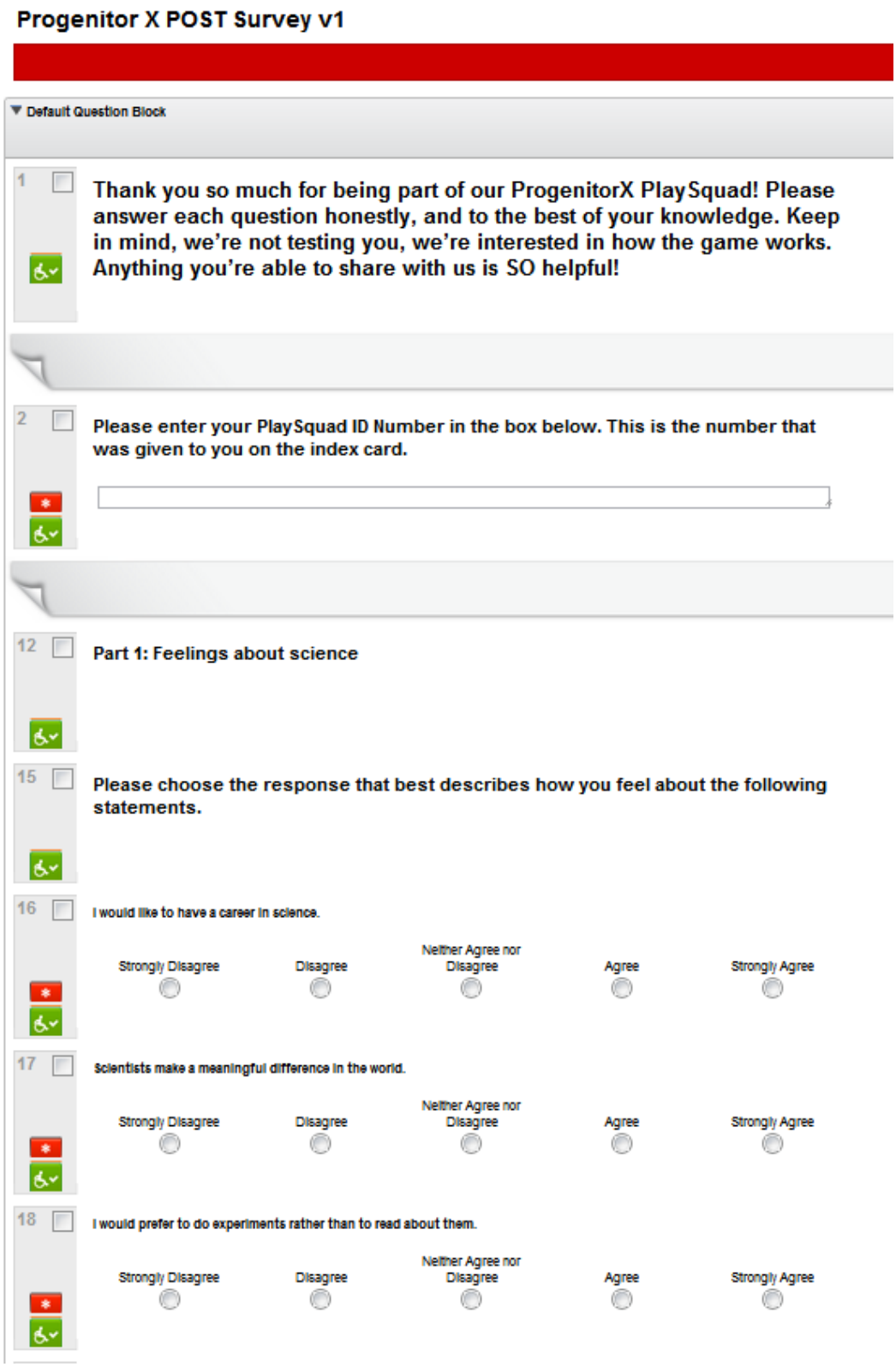
Steinkuehler, C. A. (2008). Cognition and literacy in massively multiplayer online games. In J. Coiro, M. Knobel, C. Lankshear, & D. Leu (Eds.), *Handbook of Research on New Literacies*. Mahwah NJ: Erlbaum. 611-634.

Steinkuehler, C., & Duncan, S. (2008). Scientific habits of mind in virtual worlds. *Journal of Science Education and Technology, 17*(6), 530–543.

Xu, B., & Recker, M. (2012) Teaching Analytics: A Clustering and Triangulation Study of Digital Library User Data. *Educational Technology & Society Journal, 15(3)*, 103-115.

Zapata-Rivera, D. & Bauer, M. (2011) Exploring the Role of Games in Educational Assessment. In Mayrath, M, Clarke-Midura, J., Robinson, D. and Shraw, G. *Technology-Based Assessments for 21st Century Skills: Theoretical and Practical Implications from Modern Research*. Pages 147–169. Information Age Publishing.

Appendix



Progenitor X POST Survey v1

▼ Default Question Block

1 Thank you so much for being part of our ProgenitorX PlaySquad! Please answer each question honestly, and to the best of your knowledge. Keep in mind, we're not testing you, we're interested in how the game works. Anything you're able to share with us is SO helpful!

2 Please enter your PlaySquad ID Number in the box below. This is the number that was given to you on the index card.

12 **Part 1: Feelings about science**

15 Please choose the response that best describes how you feel about the following statements.

16 I would like to have a career in science.

Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree
○	○	○	○	○

17 Scientists make a meaningful difference in the world.

Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree
○	○	○	○	○

18 I would prefer to do experiments rather than to read about them.

Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree
○	○	○	○	○

27

Part 2: Science

Please choose the answer you think is correct for each of the following q worry if you don't know the answer. Just pick the one you think is best.

28

Where can fibroblasts come from?

- They can come from plants.
- They can be artificially created in a lab.
- They can come from adult humans.
- They come from large rocks.

29

Where can iPS cells come from?

- They can come from treated fibroblasts.
- They can be made from plastic.
- They can grow from the ground.
- They can be made by mixing chemical compounds.

30

What does the growth factor do to iPS cells?

- Destroys all of them.
- Turns them into special cells that can be used to make human tissue.
- Turns them into plant cells.
- Changes them into chemicals for creating robot parts.

31

What kind of cells can iPS cells turn into?

- Roboderm, mesoderm, and ectoderm
- Mesoderm, ectoderm, and plasmaderm
- Roboderm, endoderm, and mesoderm
- Endoderm, mesoderm, and ectoderm

32

What are two steps used in transforming fibroblasts to iPS cells?

- Shocking cells with electroporation, then exposing them to the zombie virus
- Eliminating cells with bacteria, then collecting them with a micropipette
- Shocking cells with electroporation, then collecting them with a micropipette
- Collecting cells with a micropipette and then evaporating them